
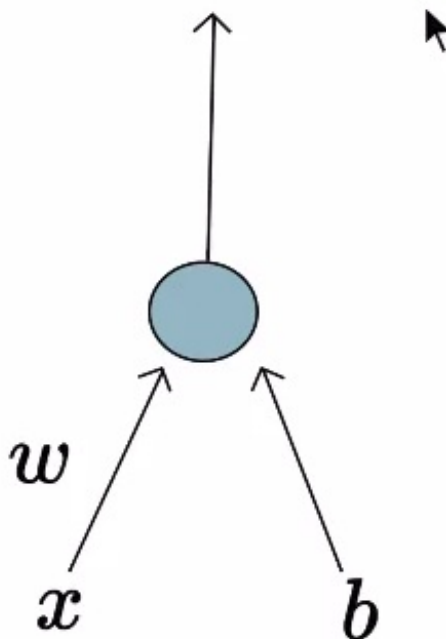
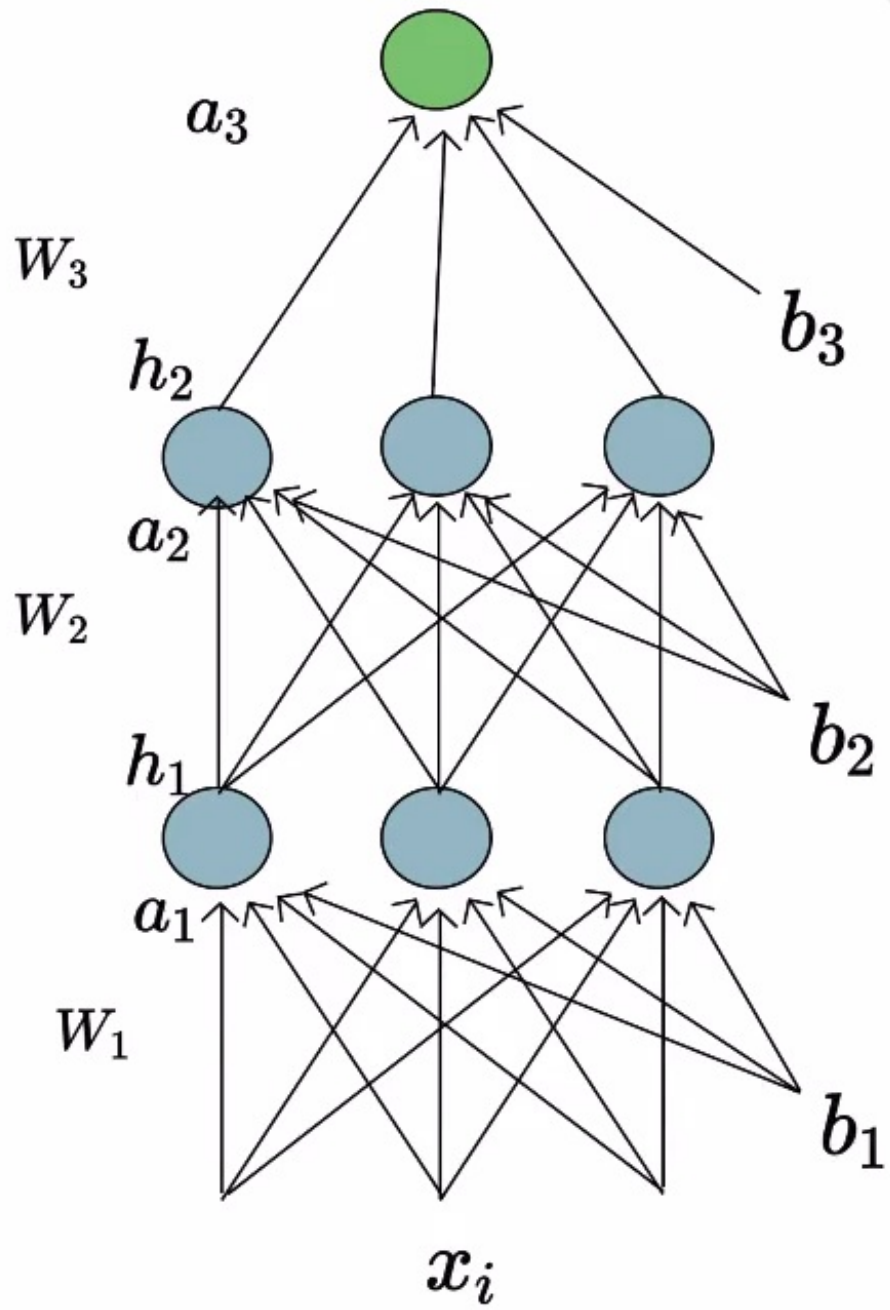


Back Propagation

 medium.com/@manveetdn/back-propagation-5b1f8bb2be94

Disclaimer: This is notes on “Back Propagation” Lesson (PadhAI [onefourthlabs](#) course “A First Course on Deep Learning”)





Before and after feed forward neural network.

The algorithm

Initialise w, b

Iterate over data:

compute \hat{y}

compute $\mathcal{L}(w, b)$

$$w_{111} = w_{111} - \eta \Delta w_{111}$$

$$w_{112} = w_{112} - \eta \Delta w_{112}$$

....

$$w_{313} = w_{313} - \eta \Delta w_{313}$$

till satisfied

The gradient descent Algorithm

1.Initialise: $w_{111}, w_{112}, \dots, w_{313}, b_i, b_i, b_i$ randomly

2.Iterate over data

a. Compute \hat{y}

b. Compute $L(w, b)$ Cross-entropy loss function

1. $w_{111} = w_{111} - \eta \Delta w_{111}$

2. $w_{112} = w_{112} - \eta \Delta w_{111}$

.....

$$3. w_{313} = w_{313} - \eta \Delta^{w_{313}}$$

$$c. b_i = b_i + \eta \Delta^{b_i}$$

Pytorch and Tensorflow have many functions to compute derivatives dw and db

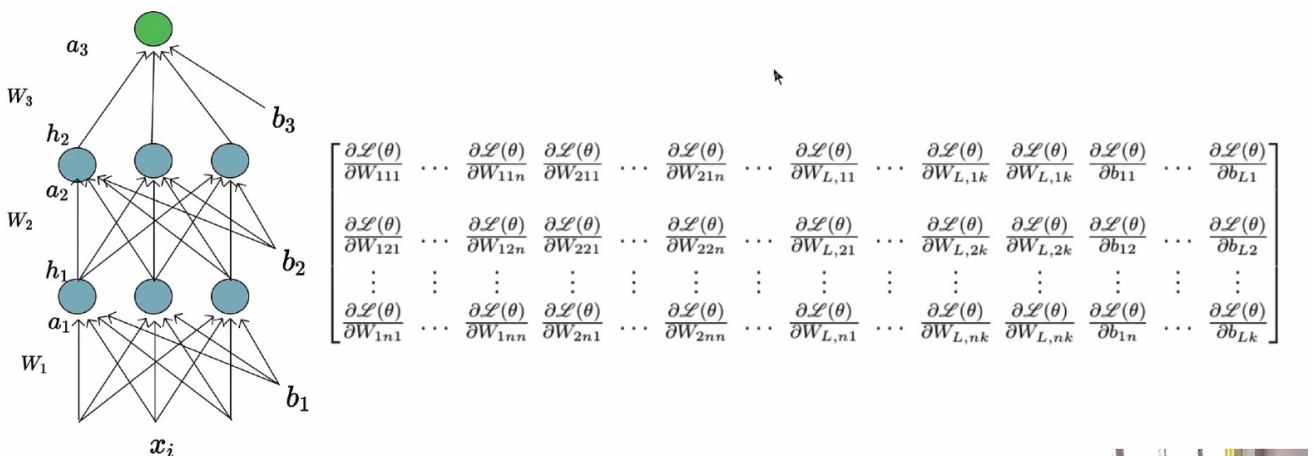
3. Till satisfied:

Number of epochs is reached (ie 1000 passes/epochs)

Continue till Loss < ϵ (some threshold).

here, we will collect the how we compute the **derivative loss function with derivative respect to 'w' and with respect to 'b'** , these are dl/dw and dl/db respectively.

Now the above part in of the Deep Neural Networks. Now , we need to compute derivative the loss function with respect to each on, that is w_{111} , w_{112} ,..... and so on, Like with all the parameters in the model as shown below.



Computing the derivative of loss functions with respect to w and b , taking help of all the parameter in the network from w , to b .

We need to compute all these derivatives and after that we can apply the gradient descent algorithm applying to update the weights.

Basic derivatives(Recap):

$$\frac{de^x}{dx} = e^x$$

$$\frac{dx^2}{dx} = 2x$$

$$\frac{d(1/x)}{dx} = -\frac{1}{x^2}$$

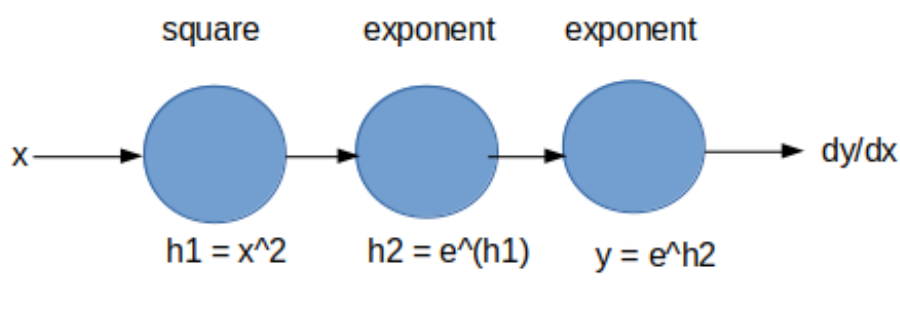
$$\frac{de^{x^2}}{dx} = \frac{de^{x^2}}{dx^2} \cdot \frac{dx^2}{dx} = \frac{de^z}{dz} \cdot \frac{dx^2}{dx} = (e^z) \cdot (2x) = (e^{x^2}) \cdot (2x) = 2xe^{x^2}$$

these are basic derivative we need to know before going further.

After all this, we need to compute other values using all these terminologies and process.

What will the value of :

$$\frac{de^{e^{x^2}}}{dx} = ?$$



This is the process followed.

Here we pass x to a square function which results in **$h1 = x^2$** and then we will pass the $h1$ to an exponent function which results in **$h2 = e^{h1}$** and after that we will pass that to another function **$y = e^{h2} = e^{(e^{x^2})}$** . Like that we find the derivative of $e^{h2} = e^{(e^{x^2})}$ as shown below.

We can write the given from like this. So, from this we will proceed further.

$$\frac{de^{e^{x^2}}}{dx} = \frac{dy}{dh2} \frac{dh2}{dh1} \frac{dh1}{dx}$$

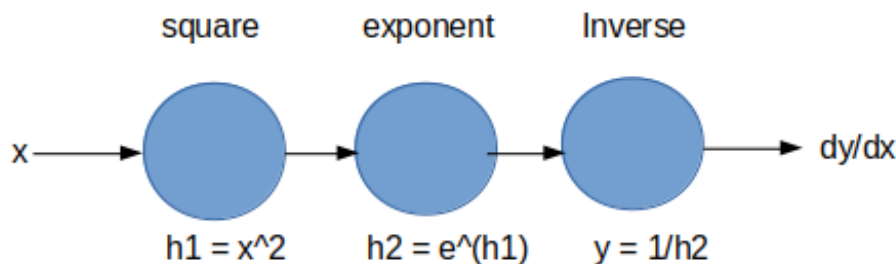
$$\frac{de^{x^2}}{dx} = \frac{de^{x^2}}{de^{x^2}} \cdot \frac{de^{x^2}}{dx} = \frac{de^z}{dz} \cdot \frac{de^{x^2}}{dx} = (e^z) \cdot (2xe^{x^2}) = (e^{e^{x^2}}) \cdot (2xe^{x^2}) = 2xe^{x^2} e^{e^{x^2}}$$

like this we find the derivative.

Similarly, we will find derivative of other

In the same way as below first we will pass x to a square function which results in $h1 = x^2$. Then, we pass $h1$ to an exponential function which results in $h2 = e^{h1}$. Then, we pass it to an inverse function we get, $y = 1/h2$.

$$\frac{d(1/e^{x^2})}{dx} = ?$$



this is the process followed.

Then we apply the chain rule and find the final value.

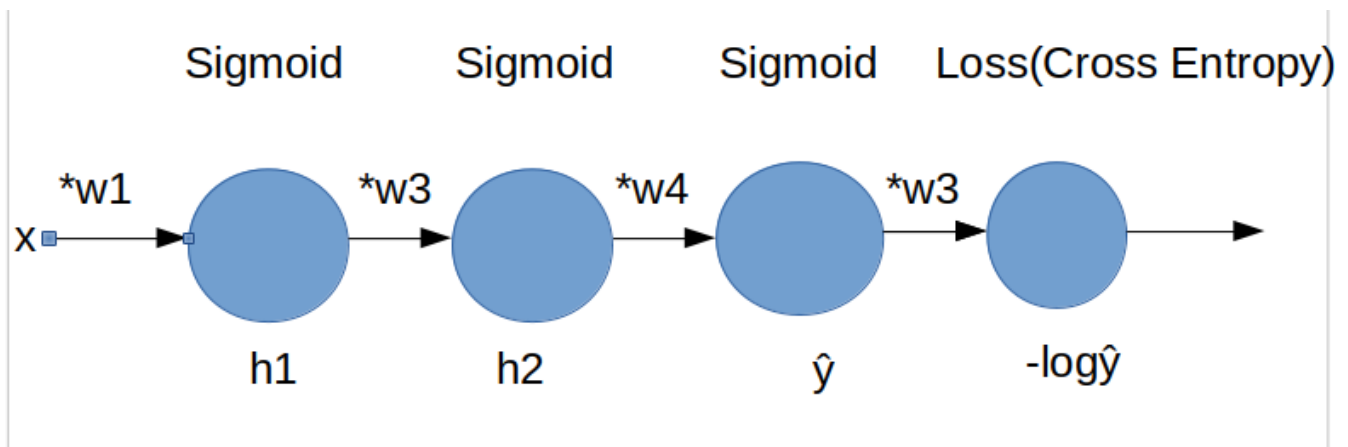
We can write the given like this and we will further proceed to compute using chain rule.

$$\frac{d(\frac{1}{e^{x^2}})}{dx} = \frac{dy}{dh2} \frac{dh2}{dh1} \frac{dh1}{dx}$$

$$\frac{d(1/e^{x^2})}{dx} = \frac{d(1/e^{x^2})}{de^{x^2}} \cdot \frac{de^{x^2}}{dx} = \frac{d(1/z)}{dz} \cdot \frac{de^{x^2}}{dx} = \left(\frac{-1}{(z)^2}\right) \cdot (2xe^{x^2}) = \left(\frac{-1}{(e^{x^2})^2}\right) \cdot (2xe^{x^2})$$

Applying chain rule and we find final output.

Why do we care about calculus and derivatives in this deep Learning??



Taking one more example.

In this example, we will take an input x in the then multiply with w_1 and we will pass that thorough a sigmoid function and get h_1 and with h_1 . We will multiply that with w_2 and after passing through the sigmoid function, we will get h_2 . Then we will pass that through sigmoid function multiplying with w_3 and we will get the final \hat{y} . **To that \hat{y} we will apply the loss function** and compute the loss.

Nowm,we can say that loss function is a function of 4 parameters in this case namely **x, w_1, w_2, w_3** .

Loss = $f(x, w_1, w_2, w_3)$.

Now the gradient, we want the derivative of the loss function with rspect to the various weights (**$\partial L / \partial w_i$**).

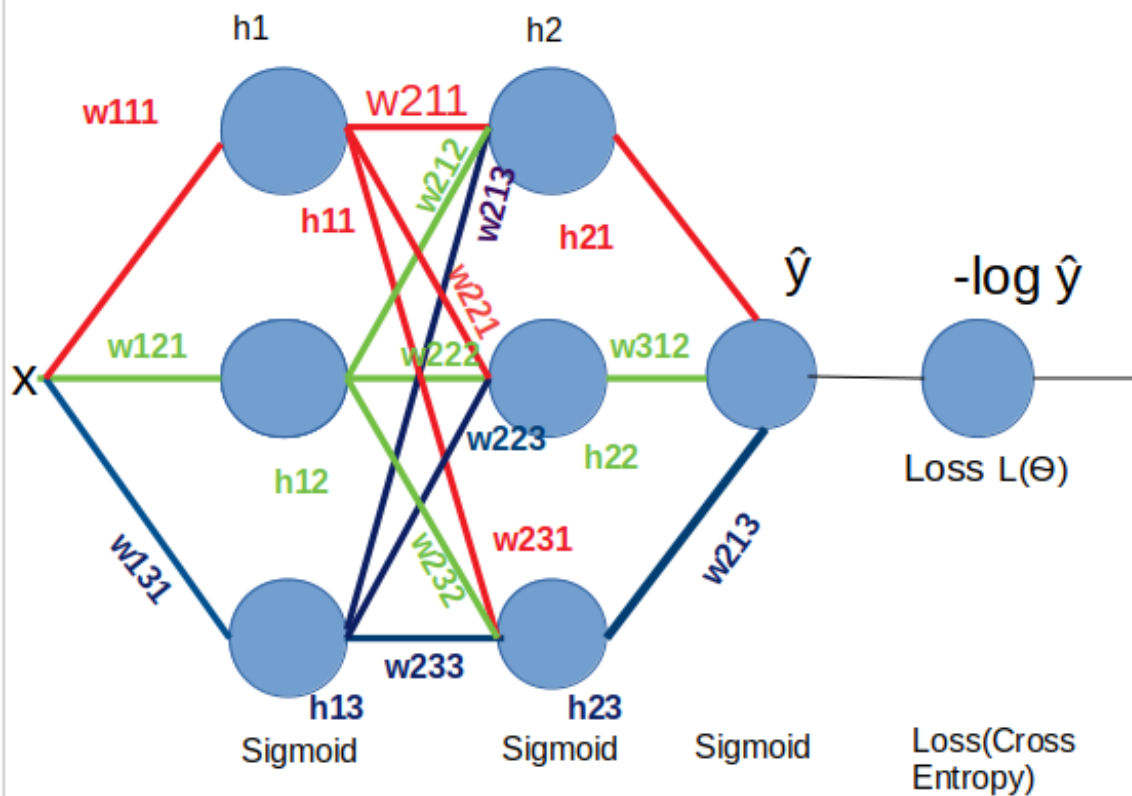
the derivative with respect to w .

Forward propagation: Here, computation happens from input layer to the output layer that is called **forward propagation**.

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial h_2} \frac{\partial h_2}{\partial w_2}$$

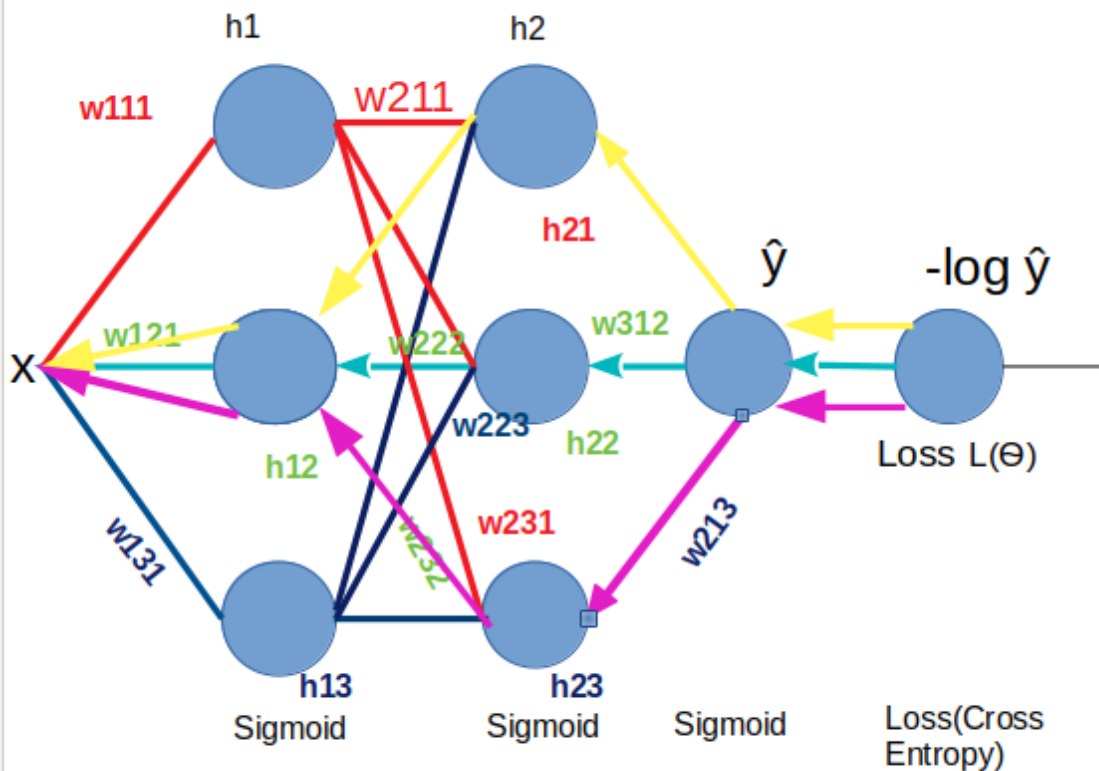
Back Propagation: Derivative calculation happens backwards from the output layer to the input, that is called as **back propagation**.

Chain rule across multiple paths:



Deep Network.

1. In the shallow Neural Network from the previous example, we apply the chain rule along a straight path. However, in a more practical Neural Network as shown above, the chain rule needs to be applied across multiple parallel paths in order to find a particular gradient.
2. For example, to calculate $\partial L / \partial w_{121}$ we need to operate along 3 different paths.

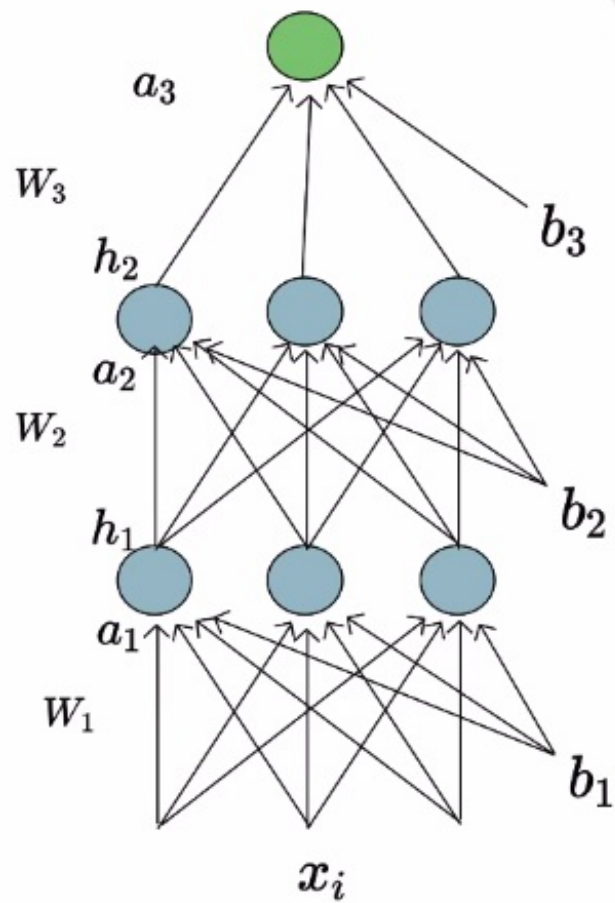


Summing up the derivatives across the three paths (blue, red and pink) will give us the required derivative $\partial L / \partial w_{121}$. This scales across as many paths as there are in the neural network. Here, these are not regular derivatives but partial derivatives.

Applying chain rule in a neural network.

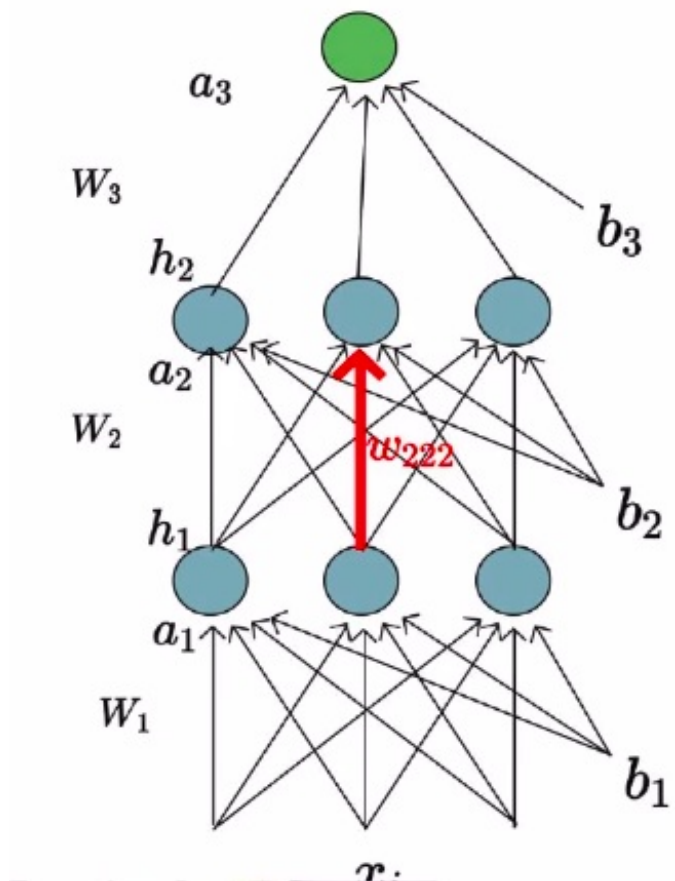
How many derivatives do we need to compute and how do we compute them?

Let's concentrate on the w_{222} cell now of the following neural network below.



the deep neural network and we focus on w_{22} which is highlighted with colour red in the second image.

The update rule which we are going to use here is



$$(w_{222})_{t+1} = (w_{222})_t - \eta * \left(\frac{\partial L}{\partial w_{222}} \right)$$

the update rule or formula.

then we will focused on w_{22} so we need to compute the partial derivative loss function with respect to w_{22} so we get the folleing as below if we want to compute that.

$$\frac{\partial L}{\partial w_{222}} = \left(\frac{\partial L}{\partial a_{22}} \right) \cdot \left(\frac{\partial a_{22}}{\partial w_{222}} \right)$$

the partial derivative of loss function with respect to w_{22} .

we calculate the partial derivative as follows

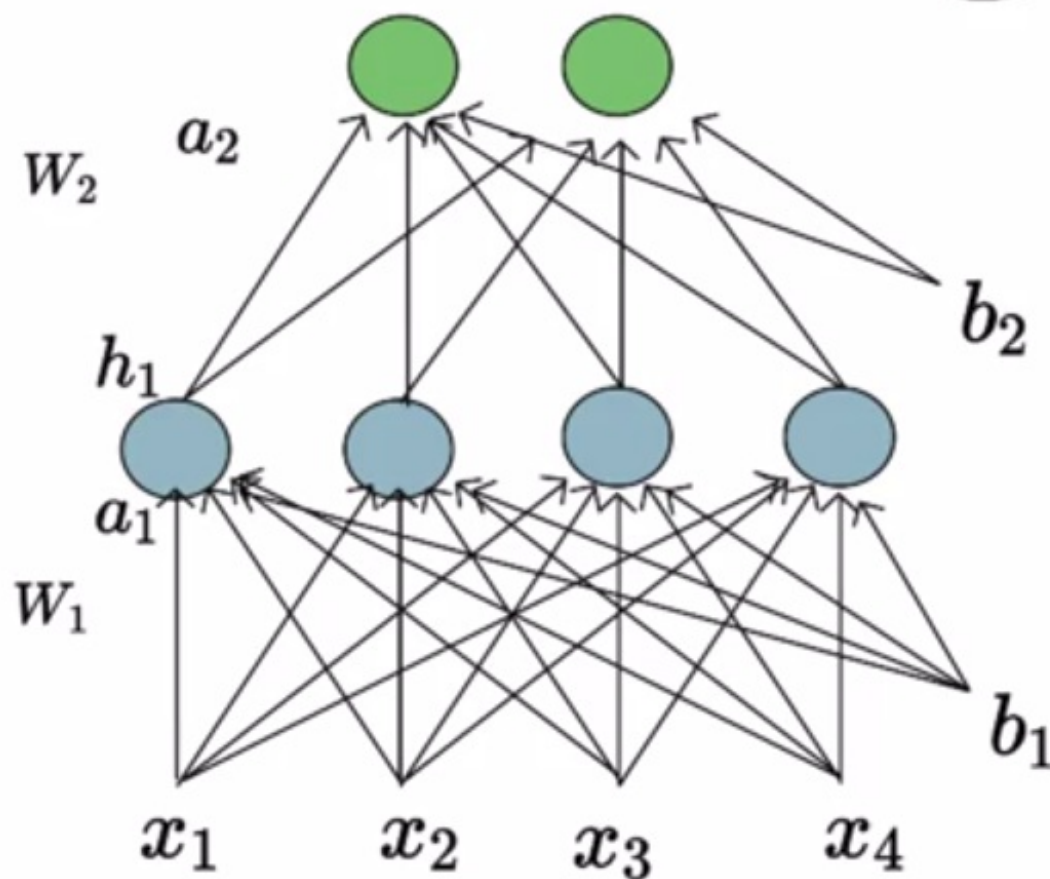
$$\begin{aligned} \frac{\partial L}{\partial w_{222}} &= \left(\frac{\partial L}{\partial a_{22}} \right) \cdot \left(\frac{\partial a_{22}}{\partial w_{222}} \right) \\ &= \left(\frac{\partial L}{\partial h_{22}} \right) \cdot \left(\frac{\partial h_{22}}{\partial a_{22}} \right) \cdot \left(\frac{\partial a_{22}}{\partial w_{222}} \right) \\ &= \left(\frac{\partial L}{\partial a_{31}} \right) \cdot \left(\frac{\partial a_{31}}{\partial h_{22}} \right) \cdot \left(\frac{\partial h_{22}}{\partial a_{22}} \right) \cdot \left(\frac{\partial a_{22}}{\partial w_{222}} \right) \\ &= \left(\frac{\partial L}{\partial \hat{y}} \right) \cdot \left(\frac{\partial \hat{y}}{\partial a_{31}} \right) \cdot \left(\frac{\partial a_{31}}{\partial h_{22}} \right) \cdot \left(\frac{\partial h_{22}}{\partial a_{22}} \right) \cdot \left(\frac{\partial a_{22}}{\partial w_{222}} \right) \end{aligned}$$

computing the partial derivative.

Thus, by breaking the partial derivative into all the subdivisions along that path and multiplying it, we will get the desired solution.

Partial derivative w.r.t a

Part 1:



Here we instantiated a neural network using this we will take one particular weight and we will find the partial derivative of that.

Here in the deep neural network, in the first layer we have **4 neurons in the input layer** those are **x_1, x_2, x_3, x_4** respectively. It also has **4 neurons in the first hidden layer** which are indicated by blue colour in the network. So our weight in the first layer will be a matrix of the order $[4 \times 4]$ and 4 biases in the first layer. Then we take the second layer we have **4 neurons in the input layer** and **2 neurons in the output layer**. So, the weight is a matrix of the order $[2 \times 4]$

$$b = \begin{bmatrix} 0 & 0 \end{bmatrix}$$

$$W_1 = \begin{bmatrix} 0.1 & 0.3 & 0.8 & -0.4 \\ -0.3 & -0.2 & 0.5 & 0.5 \\ -0.3 & 0 & 0.5 & 0.4 \\ 0.2 & 0.5 & -0.9 & 0.7 \end{bmatrix}$$

$$W_2 = \begin{bmatrix} 0.5 & 0.8 & 0.2 & 0.4 \\ 0.5 & 0.2 & 0.3 & -0.5 \end{bmatrix}$$

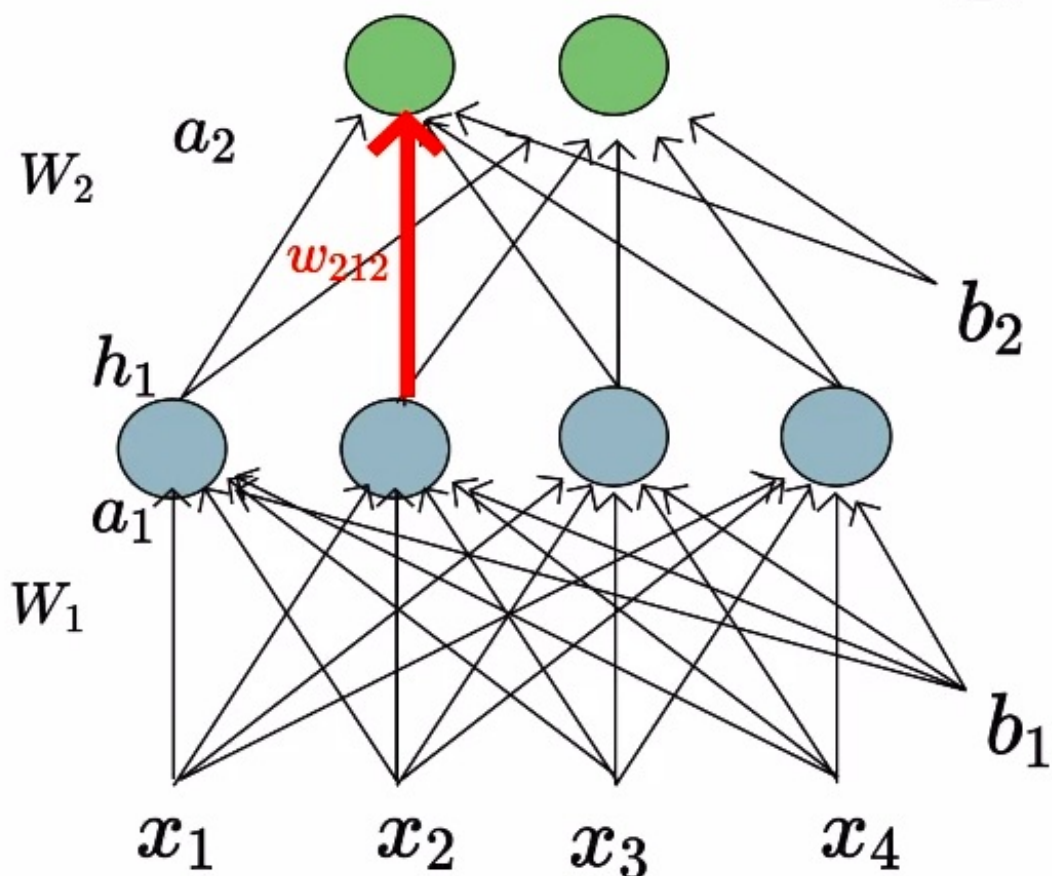
the bias matrix , 1st layer weight matrix and the 2nd layer weights matrix.

For this network we took 4 inputs and we gave 2 outputs here in this case, here the true outputs and the inputs are as shown below.

$$y = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

$$x = \begin{bmatrix} 2 & 5 & 3 & 3 \end{bmatrix}$$

True outputs values and inputs matrices.



We will take the w212 neuron here in w212 if first 2 represents that its the second layer neuron and the 1 represents that it connects 1st neuron of the last layer and last 2 represents that it connecting 2nd neuron of the second layer.

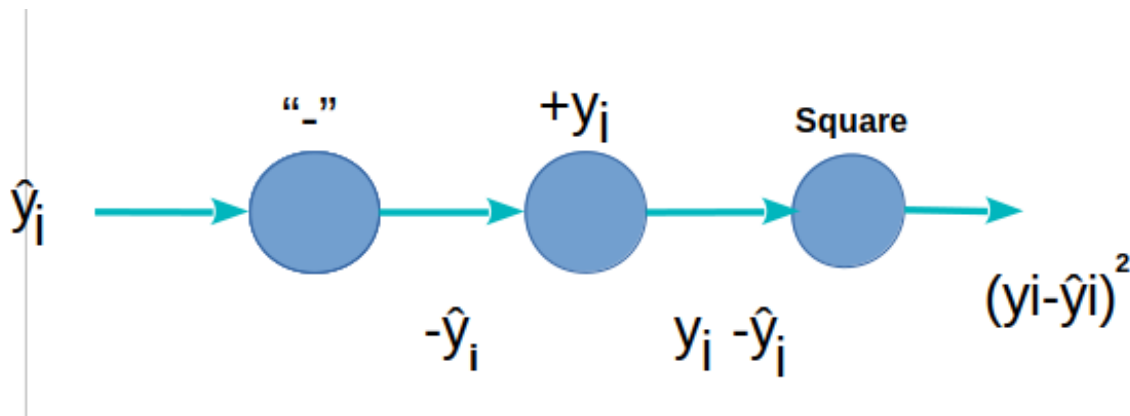
we will take that neuron and we will calculate the partial derivative of the loss function with respect to the w212

Now the partial derivative of the loss function with respect to the w212 is represented as.

$$\frac{\partial L}{\partial w_{212}} = \left(\frac{\partial L}{\partial a_{21}} \right) \cdot \left(\frac{\partial a_{21}}{\partial w_{212}} \right) = \left(\frac{\partial L}{\partial \hat{y}_1} \right) \cdot \left(\frac{\partial \hat{y}_1}{\partial a_{21}} \right) \cdot \left(\frac{\partial a_{21}}{\partial w_{212}} \right)$$

partial derivative of loss function with respect to w212.

If we consider the square error loss then we will proceed like the following



this is the process thing we follow while applying square error loss.

Computing the derivative of loss function with respective \hat{y}_i (the first term):

Here, in the above case y_2 terms get cancelled because in this its partial derivative w.r.t y_1 so those terms get cancelled.

$$\frac{\partial L}{\partial \hat{y}_1} = \sum_{i=1}^2 (y_i - \hat{y}_i)^2$$

$$\frac{\partial L}{\partial \hat{y}_1} = \frac{\partial}{\partial y_1} [(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2]$$

Like this we get the first derivative value.

Part 2:

$$\frac{\partial L}{\partial \hat{y}_1} = -2(y_1 - \hat{y}_1)$$

Now computing the second derivative of \hat{y}_i
w.r.t a_{21} :

soft max on the a_{21}

Here, \hat{y}_i is the soft max function of on the a_{21} .

$$\hat{y}_1 = \left(\frac{e^{a_{21}}}{e^{a_{21}} + e^{a_{22}}} \right)$$

1. To make it easier to compute, multiply both numerator and denominator by $(e)^{-a_{21}}$

$$\hat{y}_1 = \left(\frac{e^{-a_{21}}}{e^{-a_{21}} + e^{a_{22}}} \right) = \frac{1}{1 + e^{-(a_{21} - a_{22})}}$$

$$\frac{\partial \hat{y}_1}{\partial a_{21}} = \frac{\partial}{\partial a_{21}} \left(\frac{1}{1 + e^{-(a_{21} - a_{22})}} \right)$$

$$\frac{\partial \hat{y}_1}{\partial a_{21}} = \left(\frac{-1}{(1 + e^{-(a_{21} - a_{22})})^2} \right) \cdot (1) \cdot (e^{-(a_{21} - a_{22})}) \cdot (-1) = \left(\frac{1}{1 + e^{-(a_{21} - a_{22})}} \right) \cdot \left(\frac{e^{-(a_{21} - a_{22})}}{1 + e^{-(a_{21} - a_{22})}} \right)$$

Like this we will compute the following.

Then rearranging the terms and rewriting the whole equation the get second derivative.

$$\frac{\partial \hat{y}_1}{\partial a_{21}} = \hat{y}_1 * (1 - \hat{y}_1)(-a_{22})$$

Now we get the second derivative value.

Part 3:

Computing the derivative of a21 w.r.t **w212**:

This is the value of a21.

$$a_{21} = w_{211}h_{11} + w_{212}h_{12} + w_{213}h_{13} + w_{214}h_{14}$$

Here, a21 is written as below the some of these 3 products. Now taking the partial derivative of a21 w.r.t w212 then,

Finally we remain with this as final results as remaining all terms will be cancelled out.

Finally we found all the three derivative values, with all theses values we start computing all the values with our sample values of input and the output.

$$\frac{\partial a_{21}}{\partial w_{212}} = h_{12}$$

$$x = \begin{bmatrix} 2 & 5 & 3 & 3 \end{bmatrix}$$

$$y = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

input and the output we have.

Output:

$$a_1 = W_1 * x + b_1 = [2.9 \quad 1.4 \quad 2.1 \quad 2.3]$$

$$h_1 = \text{sigmoid}(a_1) = [0.95 \quad 0.80 \quad 0.89 \quad 0.91]$$

$$a_2 = W_2 * h_1 + b_2 = [1.66 \quad 0.45]$$

$$\hat{y} = \text{softmax}(a_2) = [0.77 \quad 0.23]$$

Here the bias matrix is taken as zero for the sake of simplicity.

And taking one more step we will compute the three derivatives as follows:

$$\frac{\partial L}{\partial w_{212}} = \left(\frac{\partial L}{\partial a_{21}} \right) \cdot \left(\frac{\partial a_{21}}{\partial w_{212}} \right) = \left(\frac{\partial L}{\partial \hat{y}_1} \right) \cdot \left(\frac{\partial \hat{y}_1}{\partial a_{21}} \right) \cdot \left(\frac{\partial a_{21}}{\partial w_{212}} \right)$$

$$\frac{\partial L}{\partial \hat{y}_1} = -2(y_1 - \hat{y}_1) = -0.46$$

$$\frac{\partial \hat{y}_1}{\partial a_{21}} = \hat{y}_1 * (1 - \hat{y}_1)(-a_{22}) = -0.079$$

$$\frac{\partial a_{21}}{\partial w_{212}} = h_{12} = 0.80$$

$$\begin{aligned} \frac{\partial L}{\partial w_{212}} &= (-2(y_1 - \hat{y}_1)) * (\hat{y}_1(1 - \hat{y}_1)(-a_{22})) * (h_{12}) \\ &= (-0.46) * (-0.079) * (0.80) = -0.029 \end{aligned}$$

Calculating all the three derivatives.

$$w_{212} = w_{212} - \eta \left(\frac{\partial L}{\partial w_{212}} \right)$$

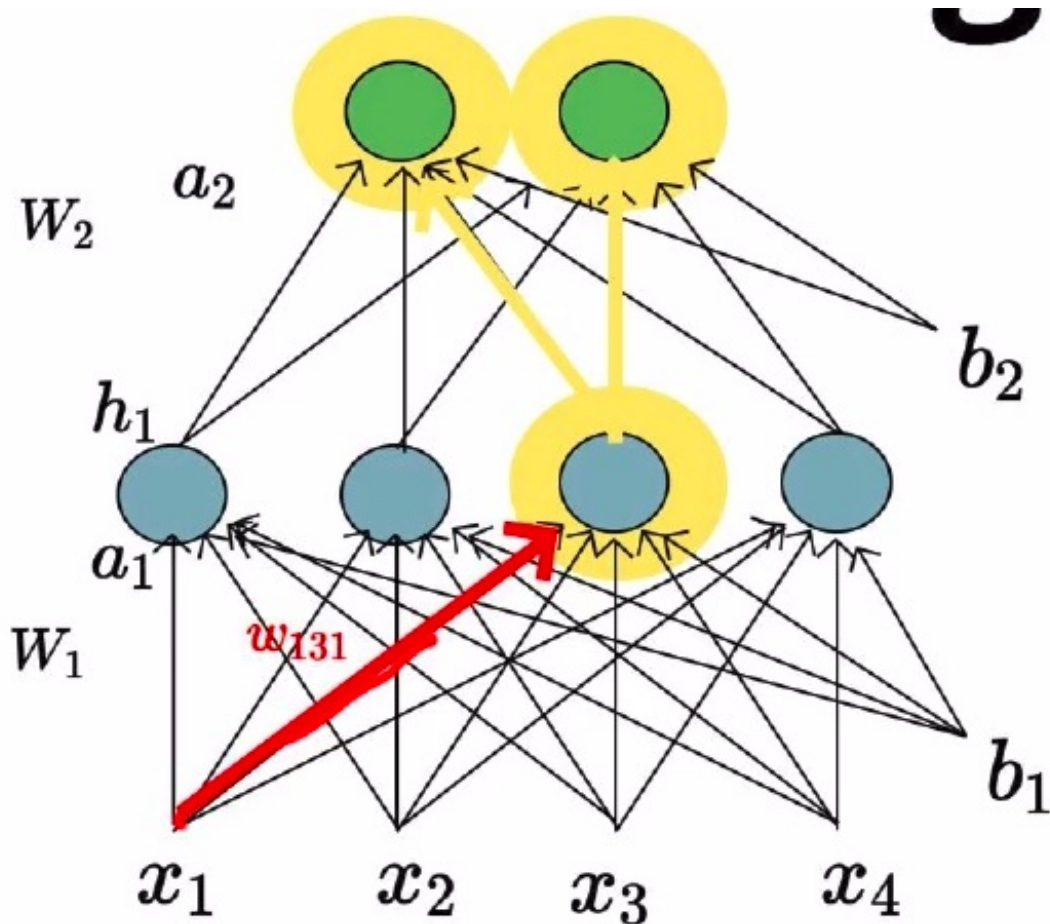
$$w_{212} = 0.8 - (1) * (-0.065)$$

$$w_{212} = 0.865$$

Like this we use the update rule

Now with all these we will get the update rule of the w_{212} with the known formula as shown above.

Computing partial derivatives w.r.t a weight when there are multiple paths:



Now in the same neural network as before if we take the w_{131} as pointed with the red colour pointer in the image beside the we need to find the derivative of the loss function with respect to the w_{131} .

Now the derivative of the loss function can be written as below.

$$\frac{\partial L}{\partial w_{131}} = \left(\frac{\partial L}{\partial a_{13}} \right) \cdot \left(\frac{\partial a_{13}}{\partial w_{131}} \right)$$

Dividing further , then we will will get the following equation.

$$\frac{\partial L}{\partial w_{131}} = \left(\frac{\partial L}{\partial a_{21}} \cdot \frac{\partial a_{21}}{\partial h_{13}} + \frac{\partial L}{\partial a_{22}} \cdot \frac{\partial a_{22}}{\partial h_{13}} \right) \cdot \left(\frac{\partial h_{13}}{\partial a_{13}} \right) \cdot \left(\frac{\partial a_{13}}{\partial w_{131}} \right)$$

Now each term in the braces can be further divided as below. and the final equation will be as follows.

$$\frac{\partial L}{\partial w_{131}} = \left(\frac{\partial L}{\partial \hat{y}_1} \cdot \frac{\partial \hat{y}_1}{\partial a_{21}} \cdot \frac{\partial a_{21}}{\partial h_{13}} + \frac{\partial L}{\partial \hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial a_{22}} \cdot \frac{\partial a_{22}}{\partial h_{13}} \right) \cdot \left(\frac{\partial h_{13}}{\partial a_{13}} \right) \cdot \left(\frac{\partial a_{13}}{\partial w_{131}} \right)$$

Here, we are using the square error loss and also some terms values are given as below.

$$\frac{\partial L}{\partial \hat{y}_1} = -2(y_1 - \hat{y}_1)$$

$$\frac{\partial L}{\partial \hat{y}_2} = -2(y_2 - \hat{y}_2)$$

$$1 \quad \frac{\partial \hat{y}_1}{\partial a_{21}} = \hat{y}_1 * (1 - \hat{y}_1) * (-a_{22})$$

$$\frac{\partial \hat{y}_2}{\partial a_{22}} = \hat{y}_2 * (1 - \hat{y}_2) * (-a_{21})$$

$$\frac{\partial a_{21}}{\partial h_{13}} = w_{213}$$

$$\frac{\partial a_{22}}{\partial h_{13}} = w_{223}$$

$$\frac{\partial h_{13}}{\partial a_{13}} = h_{13} * (1 - h_{13})$$

$$\frac{\partial a_{13}}{\partial w_{131}} = x_1 \quad \blacktriangleright$$

the are formulae to compute the values.

finally putting all this values in the final loss function derivative equation then we will finally get.

$$\frac{\partial L}{\partial w_{131}} = (-2(y_1 - \hat{y}_1) * \hat{y}_1(1 - \hat{y}_1) * w_{213} + -2(y_2 - \hat{y}_2) * \hat{y}_2(1 - \hat{y}_2) * w_{223}) * h_{13}(1 - h_{13}) * x_1$$

Now computing all the values one by one will finally end up like this.

$$\frac{\partial L}{\partial \hat{y}_1} = -2(y_1 - \hat{y}_1) = -0.46$$

$$\frac{\partial L}{\partial \hat{y}_2} = -2(y_2 - \hat{y}_2) = 0.46$$

$$\begin{aligned} \frac{\partial \hat{y}_1}{\partial a_{21}} &= \hat{y}_1 * (1 - \hat{y}_1) * (-a_{22}) \\ &= -0.079 \end{aligned}$$

$$\begin{aligned} \frac{\partial \hat{y}_2}{\partial a_{22}} &= \hat{y}_2 * (1 - \hat{y}_2) * (-a_{21}) \\ &= -0.293 \end{aligned}$$

$$\frac{\partial a_{21}}{\partial h_{13}} = w_{213} = 0.20$$

$$\frac{\partial a_{22}}{\partial h_{13}} = w_{223} = 0.30$$

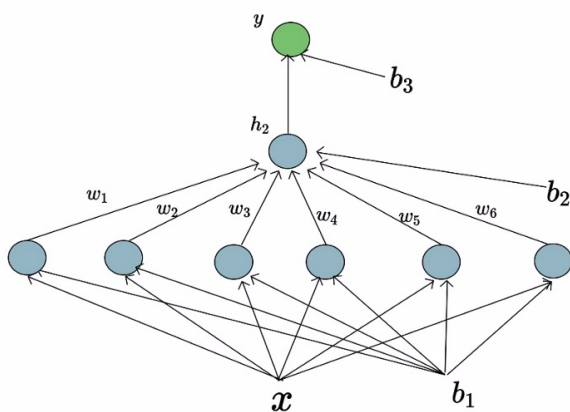
$$\frac{\partial h_{13}}{\partial a_{13}} = h_{13} * (1 - h_{13}) = 0.0979$$

$$\frac{\partial a_{13}}{\partial w_{131}} = x_1 = 2$$

Then, finally the partial derivative of loss function w.r.t w_{131} will evaluate to the following values.

$$\begin{aligned} \frac{\partial L}{\partial w_{131}} &= (-2(y_1 - \hat{y}_1) * \hat{y}_1(1 - \hat{y}_1) * w_{213} + -2(y_2 - \hat{y}_2) * \hat{y}_2(1 - \hat{y}_2) * w_{223}) * h_{13}(1 - h_{13}) * x_1 \\ &= (-0.46 * -0.079 * 0.20 + 0.46 * -0.293 * 0.3) * 0.0979 * 2 = -6.4 \times 10^{-3} \end{aligned}$$

Final Take away:



No matter how complex the function, we can always compute the derivative wrt any variable using the chain rule

We can reuse a lot of work by starting backwards and computing simpler elements in the chain

This is all about the Back propagation and chain rule applying to find the partial derivatives.

This is a small try ,uploading the notes . I believe in **"Sharing knowledge is that best way of developing skills"**.Comments will be appreciated. Even small edits can be suggested.

Each Applause will be a great encouragement.

Do follow my medium for more updates.....